PCT WELTORGANISATION FÜR GEISTIGES EIGENTUM
Internationales Büro

INTERNATIONALE ANMELDUNG VERÖFFENTLICHT NACH DEM VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS (PCT)

(51) Internationale Patentklassifikation 7 : G06K 17/60	A2	(11) Internationale Veröffentlichungsnummer: **WO 00/17811**
		(43) Internationales Veröffentlichungsdatum: 30. März 2000 (30.03.00)

(21) Internationales Aktenzeichen: PCT/DE99/02846

(22) Internationales Anmeldedatum: 8. September 1999 (08.09.99)

(30) Prioritätsdaten:
198 43 620.3 23. September 1998 (23.09.98) DE

(71) Anmelder *(für alle Bestimmungsstaaten ausser US)*: SIEMENS AKTIENGESELLSCHAFT [DE/DE]; Wittelsbacherplatz 2, D–80333 München (DE).

(72) Erfinder; und
(75) Erfinder/Anmelder *(nur für US)*: NEUNEIER, Ralf [DE/DE]; Gravelottestr. 3, D–81667 München (DE). MIHATSCH, Oliver [DE/DE]; Schulstr. 31, D–80634 München (DE).

(74) Gemeinsamer Vertreter: SIEMENS AKTIENGESELLSCHAFT; Postfach 22 16 34, D–80506 München (DE).

(81) Bestimmungsstaaten: JP, US, europäisches Patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Veröffentlicht
Ohne internationalen Recherchenbericht und erneut zu veröffentlichen nach Erhalt des Berichts.

(54) Title: METHOD AND CONFIGURATION FOR DETERMINING A SEQUENCE OF ACTIONS FOR A SYSTEM WHICH COMPRISES STATUSES, WHEREBY A STATUS TRANSITION ENSUES BETWEEN TWO STATUSES AS A RESULT OF AN ACTION

(54) Bezeichnung: VERFAHREN UND ANORDNUNG ZUR ERMITTLUNG EINER FOLGE VON AKTIONEN FÜR EIN SYSTEM, WELCHES ZUSTÄNDE AUFWEIST, WOBEI EIN ZUSTANDSÜBERGANG ZWISCHEN ZWEI ZUSTÄNDEN AUFGRUND EINER AKTION ERFOLGT

(57) Abstract

The determination of a sequence of actions ensues in such a way that a sequence of statuses resulting from the sequence of actions is optimized with regard to a predetermined optimization function. The optimization function includes a variable parameter with which a risk can be set. Said risk comprises the resulting sequence of statuses with regard to a predetermined status of the system.

(57) Zusammenfassung

Die Ermittlung der Folge von Aktionen erfolgt derart, daß eine aus der Folge von Aktionen resultierende Folge von Zuständen hinsichtlich einer vorgegebenen Optimierungsfunktion optimiert ist. Die Optimierungsfunktion enthält einen variablen Parameter, mit dem ein Risiko einstellbar ist, welches Risiko die resultierende Folge von Zuständen hinsichtlich eines vorgegebenen Zustandes des Systems aufweist.

LEDIGLICH ZUR INFORMATION

Codes zur Identifizierung von PCT-Vertragsstaaten auf den Kopfbögen der Schriften, die internationale Anmeldungen gemäss dem PCT veröffentlichen.

AL	Albanien	ES	Spanien	LS	Lesotho	SI	Slowenien
AM	Armenien	FI	Finnland	LT	Litauen	SK	Slowakei
AT	Österreich	FR	Frankreich	LU	Luxemburg	SN	Senegal
AU	Australien	GA	Gabun	LV	Lettland	SZ	Swasiland
AZ	Aserbaidschan	GB	Vereinigtes Königreich	MC	Monaco	TD	Tschad
BA	Bosnien-Herzegowina	GE	Georgien	MD	Republik Moldau	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagaskar	TJ	Tadschikistan
BE	Belgien	GN	Guinea	MK	Die ehemalige jugoslawische Republik Mazedonien	TM	Turkmenistan
BF	Burkina Faso	GR	Griechenland			TR	Türkei
BG	Bulgarien	HU	Ungarn	ML	Mali	TT	Trinidad und Tobago
BJ	Benin	IE	Irland	MN	Mongolei	UA	Ukraine
BR	Brasilien	IL	Israel	MR	Mauretanien	UG	Uganda
BY	Belarus	IS	Island	MW	Malawi	US	Vereinigte Staaten von Amerika
CA	Kanada	IT	Italien	MX	Mexiko		
CF	Zentralafrikanische Republik	JP	Japan	NE	Niger	UZ	Usbekistan
CG	Kongo	KE	Kenia	NL	Niederlande	VN	Vietnam
CH	Schweiz	KG	Kirgisistan	NO	Norwegen	YU	Jugoslawien
CI	Côte d'Ivoire	KP	Demokratische Volksrepublik Korea	NZ	Neuseeland	ZW	Zimbabwe
CM	Kamerun			PL	Polen		
CN	China	KR	Republik Korea	PT	Portugal		
CU	Kuba	KZ	Kasachstan	RO	Rumänien		
CZ	Tschechische Republik	LC	St. Lucia	RU	Russische Föderation		
DE	Deutschland	LI	Liechtenstein	SD	Sudan		
DK	Dänemark	LK	Sri Lanka	SE	Schweden		
EE	Estland	LR	Liberia	SG	Singapur		

Beschreibung

Verfahren und Anordnung zur Ermittlung einer Folge von Aktionen für ein System, welches Zustände aufweist, wobei ein Zustandsübergang zwischen zwei Zuständen aufgrund einer Aktion erfolgt

Die Erfindung betrifft ein Verfahren sowie eine Anordnung zur Ermittlung einer Folge von Aktionen für ein System, welches Zustände aufweist, wobei ein Zustandsübergang zwischen zwei Zuständen aufgrund einer Aktion erfolgt.

Ein solches Verfahren und eine solche Anordnung sind aus [1] bekannt.

In [1] ist als Beispiel für ein solches System, welches Zustände aufweist, ein Finanzmarkt beschrieben.

Das System wird als ein Markov-Entscheidungsproblem beschrieben (Markov-Decision-Problem, MDP). Ein System, welches als Markov-Entscheidungsproblem beschrieben werden kann, ist in seiner Struktur in **Fig.2** dargestellt.

Zu einem Zeitpunkt t befindet sich das System 201 in einem Zustand x_t. Der Zustand x_t ist für einen Beobachter des Systems beobachtbar. Aufgrund einer Aktion a_t aus einer Menge in dem Zustand x_t möglicher Aktionen, $a_t \in A(x_t)$ geht das System mit einer gewissen Wahrscheinlichkeit in einen Folgezustand x_{t+1} zu einem Folgezeitpunkt t+1 über.

Dies ist durch eine Schleife in **Fig.2** symbolisch dargestellt. Ein Beobachter 200 nimmt beobachtbare Größen über den Zustand x_t wahr 202 und trifft eine Entscheidung über eine Aktion 203, mit der er auf das System 201 einwirkt. Das System 201 unterliegt üblicherweise einer Störung 205.

Ferner erhält der Beobachter 200 einen Gewinn r_t 204

$$r_t = r(x_t, a_t, x_{t+1}) \in \Re, \qquad (1)$$

der von der Aktion a_t 203 und dem ursprünglichen Zustand x_t zu dem Zeitpunkt t sowie dem Folgezustand x_{t+1} des Systems zu dem Folgezeitpunkt t+1 abhängt.

Der Gewinn r_t kann einen positiven oder negativen skalaren Wert annehmen, je nachdem, ob die Entscheidung zu einer hinsichtlich eines vorgebbaren Kriteriums positiven oder negativen Systementwicklung führt, in [1] zu einer Kapitalvermehrung oder zu einem Verlust.

In einem weiteren Zeitschritt entscheidet sich der Beobachter 200 des Systems 201 aufgrund der beobachtbaren Größen 202, 204 des Folgezustandes x_{t+1} für eine neue Aktion a_{t+1} usw.

Eine Folge von

Zustand:	x_t	$\in$	X
Aktion:	a_t	$\in$	$A(x_t)$
Folgezustand:	x_{t+1}	$\in$	X
Gewinn	$r_t = r(x_t, a_t, x_{t+1})$	$\in$	$\Re$

usw. beschreibt eine Trajektorie des Systems, die durch ein Performanzkriterium, das die einzelnen Gewinne r_t über die Zeitpunkte t akkumuliert, bewertet wird. Bei einem Markov-Entscheidungsproblem wird vereinfachend angenommen, daß der Zustand x_t und die Aktion a_t alle Informationen enthalten, um eine Übergangswahrscheinlichkeit $p(x_{t+1}|\cdot)$ des Systems von dem Zustand x_t zu dem Folgezustand x_{t+1} zu beschreiben.

Formal bedeutet dies:

$$p(x_{t+1}|x_t, K, x_0, a_t, K, a_0) = p(x_{t+1}|x_t, a_t). \qquad (2)$$

Mit $p(x_{t+1}|x_t, a_t)$ wird eine Übergangswahrscheinlichkeit für den Folgezustand x_{t+1} bei gegebenem Zustand x_t und gegebener Aktion a_t bezeichnet.

Bei einem Markov-Entscheidungsproblem hängen also zukünftige Zustände des Systems 201 nicht von Zuständen und Aktionen ab, die weiter als einen Zeitschritt in der Vergangenheit liegen.

Zusammenfassend sind im weiteren die Charakteristika eines Markov-Entscheidungsproblems dargestellt:

X	Menge der möglichen Zustände des Systems, z.B. $X = \Re^m$,
$A(x_t)$	Menge der möglichen Aktionen in dem Zustand
$p(x_{t+1}\|x_t, a_t)$	x_t
$r(x_t, a_t, x_{t+1})$	Gewinn mit Erwartungswert $R(x_t, a_t)$.

Das Ziel ist es, ausgehend von beobachtbaren Größen, den im weiteren als Trainingsdaten bezeichneten Größen, eine Strategie zu ermitteln, d.h. eine Folge von Funktionen

$$\pi = \{\mu_0, \mu_1, K, \mu_T\}, \tag{3}$$

welche zu jedem Zeitpunkt t jeden Zustand in eine Handlungsvorschrift, d.h. Aktion

$$\mu_t(x_t) = a_t \tag{4}$$

abbilden.

Eine solche Strategie wird durch eine Optimierungsfunktion bewertet.

Die Optimierungsfunktion gibt den Erwartungswert, der über die Zeit akkumulierten Gewinne bei einer gegebenen Strategie π und einem Startzustand x_0 an.

Als ein Beispiel eines Verfahrens des approximativen dynamischen Programmierens ist in [1] das sogenannte Q-Lernverfahren beschrieben.

Eine optimale Bewertungsfunktion $V^*(x)$ ist definiert durch

$$V^*(x) = \max_{\pi} V^{\pi}(x) \qquad \forall x \in X \tag{5}$$

mit

$$V^{\pi}(x) = E\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, \mu_t, x_{t+1}) \middle| x_0 = x\right], \tag{6}$$

wobei mit γ ein vorgebbarer Verringerungsfaktor bezeichnet wird, der gemäß folgender Vorschrift gebildet wird:

$$\gamma = \frac{1}{1+z}, \tag{7}$$

$$z \in \Re^+. \tag{8}$$

Im Rahmen des Q-Lernverfahrens wird für jedes Paar (Zustand x_t, Aktion a_t) eine Q-Bewertungsfunktion $Q^*(x_t, a_t)$ gemäß folgender Vorschrift gebildet:

$$Q^*(x_t, a_t) := \sum_{x \in X} p(x_{t+1} | x_t, a_t) \cdot r_t + \gamma \cdot \sum_{x \in X} p(x | x_t, a_t) \cdot \max_{a \in A}\left(Q^*(x, a)\right) \tag{9}$$

Aufgrund jeweils des Tupels (x_t, x_{t+1}, a_t, r_t) werden die Q-Werte $Q^*(x,a)$ in der k+1 ten Iteration gemäß folgender Lernregel mit einer vorgegebenen Lernrate η_k gemäß folgender Vorschrift adaptiert:

$$Q_{k+1}(x_t, a_t) = (1 - \eta_k)Q_k(x_t, a_t) + \eta_k\left(r_t + \gamma \max_{a \in A}\left(Q_k(x_{t+1}, a)\right)\right). \quad (10)$$

Üblicherweise werden die sogenannten Q-Werte $Q^*(x,a)$ durch jeweils einen Funktionsapproximator, beispielsweise ein neuronales Netz oder auch einen Polynomklassifikator, mit einem Gewichtsvektor w^a, der Gewichte des Funktionsapproximators enthält, für verschiedene Aktionen a approximiert.

Unter einem Funktionsapproximator ist beispielsweise ein neuronales Netz, ein Polynomklassifikator oder auch eine Kombination eines neuronalen Netzes mit einem Polynomklassifikator zu verstehen.

Es gilt also:

$$Q^*(x, a) \approx Q\left(x; w^a\right). \quad (11)$$

Änderungen der Gewichte in dem Gewichtsvektor w^a basieren auf einer temporären Differenz d_t, die gemäß folgender Vorschrift gebildet wird:

$$d_t := r(x_t, a_t, x_{t+1}) + \gamma \max_{a \in A} Q\left(x_{t+1}; w_k^a\right) - Q\left(x_t; w_k^{a_t}\right) \quad (12)$$

Es ergibt sich für das Q-Lernverfahren unter Verwendung eines neuronalen Netzes folgende Adaptionsvorschrift für die Gewichte des neuronalen Netzes, welche Gewichte in dem Gewichtsvektor w^a enthalten sind:

$$w_{k+1}^{a_t} = w_k^{a_t} + \eta_k \cdot d_t \cdot \nabla Q\left(x_t; w_k^{a_t}\right). \qquad (13)$$

Unter Verwendung der Trainingsdaten, die als Zeitreihenwerte Informationen über vorangegangene Kursverläufe eines Finanzmarktes beschreiben, wird das neuronale Netz, welches das System Finanzmarkt, wie in [1] beschrieben, darstellt, trainiert.

Ein weiteres Verfahren der approximativen dynamischen Programmierung, das sogenannte TD(λ)-Lernverfahren, ist aus [2] bekannt und wird im Zusammenhang mit einem Ausführungsbeispiel näher erläutert.

Ferner ist aus [3] bekannt, welches Risiko mit einer Strategie π und einem Ausgangszustand x_t verbunden ist. Ein Verfahren zur Risikovermeidung ist ebenfalls aus [3] bekannt.

Bei dem aus [3] bekannten Verfahren wird folgende Optimierungsfunktion, welche auch als erweiterte Q-Funktion $\underline{Q}^{\pi}(x_t, a_t)$ bezeichnet wird, verwendet:

maximiere

$$\left(\underline{Q}^{\pi}(x_t, a_t) := r(x_t, a_t, x_{t+1}) + \inf_{\substack{x_0, x_1, K \\ p(x_0, x_1, K) > 0}} \left\{ \sum_{k=1}^{\infty} \gamma^k r(x_k, \pi(x_k), x_{k+1}) \right\} \right) \qquad (14)$$

Die erweiterte Q-Funktion $\underline{Q}^{\pi}(x_t, a_t)$ beschreibt den schlechtesten Fall, falls in dem Zustand x_t die Aktion a_t ausgeführt wird und die Strategie π daraufhin verfolgt wird.

Die Optimierungsfunktion $\underline{Q}^{\pi}(x_t, a_t)$ für

$$\underline{Q}^*(x_t, a_t) := \max_{\pi \in \Pi} \underline{Q}^{\pi}(x_t, a_t) \quad (15)$$

ist gegeben, durch folgende Vorschrift:

$$\underline{Q}^*(x_t, a_t) = \min_{\substack{x \in X \\ p(x_{t+1}|x_t, a_t) > 0}} \left(r(x_t, a_t, x) + \gamma \cdot \max_{a \in A} \underline{Q}^*(x, a) \right). \quad (16)$$

Ein erheblicher Nachteil dieser Vorgehensweise ist darin zu sehen, daß nur der schlechteste Fall im Rahmen der Strategiefindung berücksichtigt wird. Dies spiegelt jedoch die Anforderungen verschiedenster technischer Systeme nur in unzureichendem Ausmaß wieder.

Aus [4] ist es ferner bekannt, eine Zugangskontrolle für ein Kommunikationsnetz sowie das Routing innerhalb des Kommunikationsnetzes als ein Problem der dynamischen Programmierung zu formulieren.

Somit liegt der Erfindung das Problem zugrunde, ein Verfahren sowie eine Anordnung zur Ermittlung einer Folge von Aktionen für ein System anzugeben, bei dem bzw. bei der eine erhöhte Flexibilität bei der Ermittlung der Strategie erreicht wird.

Das Problem wird durch das Verfahren sowie durch die Anordnung gemäß den Merkmalen der unabhängigen Patentansprüche gelöst.

Bei einem Verfahren zur rechnergestützten Ermittlung einer Folge von Aktionen für ein System, welches Zustände aufweist, wobei ein Zustandsübergang zwischen zwei Zuständen aufgrund einer Aktion erfolgt, erfolgt die Ermittlung der Folge von Aktionen derart, daß eine aus der Folge von Aktionen resultierende Folge von Zuständen hinsichtlich einer vorgegebenen Optimierungsfunktion optimiert ist, wobei die Optimierungs-

funktion einen variablen Parameter enthält, mit dem ein Risiko, welches die resultierende Folge von Zuständen hinsichtlich eines vorgegebenen Zustandes des Systems aufweist, einstellbar ist.

Eine Anordnung zur Ermittlung einer Folge von Aktionen für ein System, welches Zustände aufweist, wobei ein Zustandsübergang zwischen zwei Zuständen aufgrund einer Aktion erfolgt, weist einen Prozessor auf, der derart eingerichtet ist, daß die Ermittlung der Folge von Aktionen derart erfolgen kann, daß eine aus der Folge von Aktionen resultierende Folge von Zuständen hinsichtlich einer vorgegebenen Optimierungsfunktion optimiert ist, wobei die Optimierungsfunktion einen variablen Parameter enthält, mit dem ein Risiko, welches die resultierende Folge von Zuständen hinsichtlich eines vorgegebenen Zustandes des Systems aufweist, einstellbar ist.

Durch die Erfindung wird es erstmals möglich, in frei vorgebbarer Genauigkeit im Rahmen einer Strategiefindung für eine möglichen Regelung oder Steuerung, allgemein einer Beeinflussung des Systems, ein Verfahren zur Ermittlung einer Folge von Aktionen anzugeben.

Bevorzugte Weiterbildungen der Erfindung ergeben sich aus den abhängigen Ansprüchen.

Die im weiteren beschriebenen Weiterbildungen gelten sowohl für das Verfahren als auch die Anordnung, wobei bei der Weiterbildung der Anordnung jeweils der Prozessor derart eingerichtet ist, daß die Weiterbildung realisierbar ist.

In einer bevorzugten Ausgestaltung wird zur Ermittlung ein Verfahren des approximativen dynamischen Programmierens eingesetzt, beispielsweise ein auf dem Q-Lernen basierendes Verfahren oder auch ein auf dem TD(λ)-Lernen basierendes Verfahren.

Im Rahmen des Q-Lernens wird bevorzugt die Optimierungsfunktion OFQ gemäß folgender Vorschrift gebildet:

$$OFQ = Q\left(x; w^a\right),$$

wobei mit

- x ein Zustand in einem Zustandsraum X,
- a eine Aktion aus einem Aktionsraum A,
- w^a die zur Aktion a gehörigen Gewichte eines Funktionsapproximators

bezeichnet wird/werden.

Im Rahmen des Q-Lernens wird zur Ermittlung der optimalen Gewichte w^a des Funktionsapproximators folgender Adaptionsschritt ausgeführt:

$$w_{t+1}^{a_t} = w_t^{a_t} + \eta_t \cdot \aleph^{\kappa}\left(d_t\right) \cdot \nabla Q\left(x_t; w_t^{a_t}\right)$$

mit der Abkürzung

$$d_t = r\left(x_t, a_t, x_{t+1}\right) + \gamma \max_{a \in A} Q\left(x_{t+1}, w_t^a\right) - Q\left(x_t, w_t^{a_t}\right)$$

wobei mit

- x_t, x_{t+1} jeweils ein Zustand in dem Zustandsraum X,
- a_t eine Aktion aus einem Aktionsraum A,
- γ ein vorgebbarer Verringerungsfaktor,
- $w_t^{a_t}$ der zur Aktion a_t gehörige Gewichtsvektor vor dem Adaptionsschritt,
- $w_{t+1}^{a_t}$ der zur Aktion a_t gehörige Gewichtsvektor nach dem Adaptionsschritt,
- η_t (t = 1, ...) eine vorgebbare Schrittweitenfolge,

- $\kappa \in [-1; 1]$ ein Risikokontrollparameter,
- $\aleph^{\kappa}$ eine Risikokontrollfunktion $\aleph^{\kappa}(\xi) = (1 - \kappa \mathrm{sign}(\xi))\xi$,
- $\nabla Q(\cdot;\cdot)$ die Ableitung des Funktionsapproximators nach seinen Gewichten,
- $r(x_t, a_t, x_{t+1})$ ein Gewinn beim Zustandsübergang von dem Zustand x_t nach dem Folgezustand x_{t+1},

bezeichnet wird/werden.

Im Rahmen des TD(λ)-Lernverfahrens wird die Optimierungsfunktion bevorzugt gemäß folgender Vorschrift gebildet:

$$OFTD = J(x; w)$$

wobei mit

- x ein Zustand in einem Zustandsraum X,
- a eine Aktion aus einem Aktionsraum A,
- w die Gewichte eines Funktionsapproximators

bezeichnet wird/werden.

Im Rahmen des TD(λ)-Lernens wird zur Ermittlung der optimalen Gewichte w des Funktionsapproximators folgender Adaptionsschritt ausgeführt:

$$w_{t+1} = w_t + \eta_t \cdot \aleph^{\kappa}(d_t) \cdot z_t$$

mit den Abkürzungen

$$d_t = r(w_t, a_t, x_{t+1}) + \gamma J(x_{t+1}; w_t) - J(x_t; w_t),$$

$$z_t = \lambda \cdot \gamma \cdot z_{t-1} + \nabla J(x_t; w_t),$$

$$z_{-1} = 0,$$

wobei mit

- x_t, x_{t+1} jeweils ein Zustand in dem Zustandsraum X,
- a_t eine Aktion aus einem Aktionsraum A,
- γ ein vorgebbarer Verringerungsfaktor,
- w_t der Gewichtsvektor vor dem Adaptionsschritt,
- w_{t+1} der Gewichtsvektor nach dem Adaptionsschritt,
- η_t (t = 1, ...) eine vorgebbare Schrittweitenfolge,
- $\kappa \in [-1; 1]$ ein Risikokontrollparameter,
- $\aleph^\kappa$ eine Risikokontrollfunktion $\aleph^\kappa(\xi) = (1 - \kappa \mathrm{sign}(\xi))\xi$,
- $\nabla J(\cdot,\cdot)$ die Ableitung des Funktionsapproximators nach seinen Gewichten,
- $r(x_t, a_t, x_{t+1})$ ein Gewinn beim Zustandsübergang von dem Zustand x_t nach dem Folgezustand x_{t+1},

bezeichnet wird/werden.

Das System ist bevorzugt ein technisches System, von dem vor der Ermittlung Meßgrößen gemessen werden, die bei der Ermittlung der Folge von Aktionen verwendet werden.

Unter Verwendung der ermittelten Folge von Aktionen kann das technische System gesteuert oder auch geregelt werden.

Bevorzugt wird das System als ein Markov-Entscheidungsproblem modelliert.

Das Verfahren oder die Anordnung werden bevorzugt in einem Verkehrsleitsystem oder in einem Kommunikationssystem eingesetzt, wobei in dem Kommunikationssystem die Folge von Aktionen zur Durchführung einer Zugangskontrolle oder eines Routings, also einer Pfadvergabe, in einem Kommunikationsnetz eingesetzt wird.

Ferner kann das System ein Finanzmarkt sein, welcher durch ein Markov-Entscheidungsproblem modelliert wird und wobei der Verlauf des Finanzmarkts, beispielsweise ein Verlauf eines

Aktienindex oder auch ein Kursverlauf eines Devisenmarktes unter Verwendung der Verfahren bzw. der Anordnung analysiert und in den Markt entsprechend der Folge ermittelter Aktionen eingegriffen werden kann.

Ausführungsbeispiele der Erfindung sind in den Figuren dargestellt und werden im weiteren näher erläutert.

Es zeigen

Figur 1 ein Ablaufdiagramm, in dem einzelne Verfahrensschritte des ersten Ausführungsbeispiels dargestellt sind;

Figur 2 eine Skizze eines Systems, welches als Markov-Entscheidungsproblem modelliert werden kann;

Figur 3 eine Skizze eines Kommunikationsnetzes, bei dem in einer Vermittlungseinheit eine Zugangskontrolle durchgeführt wird;

Figur 4 eine symbolische Skizze eines Funktionsapproximators, mit dem ein Verfahren des approximativen dynamischen Programmierens realisiert wird;

Figur 5 eine weitere Skizze von mehreren Funktionsapproximatoren, mit dem ein approximatives dynamisches Programmieren implementiert wird;

Figur 6 eine Skizze eines Verkehrsleitsystems, welches gemäß einem Ausführungsbeispiel geregelt wird.

Erstes Ausführungsbeispiel: Zugangskontrolle und Routing.

Fig.3 zeigt ein Kommunikationsnetz 300, welches eine Vielzahl von Vermittlungseinheiten 301a, 301b, ..., 301i, ... 301n aufweist, die über Verbindungen 302a, 302b, 302j, ... 302m miteinander verbunden sind.

Ferner ist ein erstes Endgerät 303 mit einer ersten Vermittlungseinheit 301a verbunden. Von dem ersten Endgerät 303 wird eine Anforderungsnachricht 304 an die erste Vermittlungseinheit 301a gesendet, mit der eine Reservierung einer vorgegebenen Bandbreite innerhalb des Kommunikationsnetzes 300 zur Übertragung von Daten (Videodaten, textuelle Daten) angefordert wird.

In der ersten Vermittlungseinheit 301a wird gemäß einer im weiteren beschriebenen Strategie ermittelt, ob die angeforderte Bandbreite in dem Kommunikationsnetz 300 auf einer angegebenen, angeforderten Verbindung verfügbar ist (Schritt 305).

Ist dies nicht der Fall, so wird die Anforderung zurückgewiesen (Schritt 306).

Ist ausreichend Bandbreite verfügbar, so wird in einem weiteren Überprüfungsschritt (Schritt 307) überprüft, ob die Bandbreite reserviert werden kann.

Ist dies nicht der Fall, so wird die Anforderung zurückgewiesen (Schritt 308).

Sonst wird von der ersten Vermittlungseinheit 301a eine Route von der ersten Vermittlungseinheit 301a über weitere Vermittlungseinheiten 301i zu einem zweiten Endgerät 309, mit dem das erste Endgerät 303 kommunizieren will, ausgewählt und es wird eine Verbindung initialisiert (Schritt 310).

Im folgenden wird von einem Kommunikationsnetz 300 ausgegangen, welches einen Satz von Vermittlungseinheiten

$$N = \{1, K, n, K, N\} \tag{17}$$

und einen Satz von physikalischen Verbindungen

$$L = \{1, K, l, K, L\}, \tag{18}$$

umfaßt, wobei eine physikalische Verbindung l eine Kapazität von B(l) Bandbreiteneinheiten aufweist.

Es sind ein Satz

$$M = \{1, K, m, K, M\} \tag{19}$$

verschiedener Diensttypen m verfügbar, wobei ein Diensttyp m durch

- einen Bandbreitenbedarf b(m),
- eine durchschnittliche Verbindungsdauer $\frac{1}{v(m)}$, und
- einen Gewinn c(m), den man dann erhält, wenn eine Verbindungsanforderung des entsprechenden Diensttyps m akzeptiert wird,

charakterisiert ist.

Der Gewinn c(m) ist gegeben durch die Menge des Geldes, die ein Netzwerkbetreiber des Kommunikationsnetzes 300 einem Teilnehmer für eine Verbindung des Diensttyps in Rechnung stellt. Anschaulich spiegelt der Gewinn c(m) unterschiedliche, von dem Netzbetreiber vorgebbare Prioritäten wider, die er mit verschiedenen Diensten assoziiert.

Eine physikalischen Verbindung l kann gleichzeitig eine beliebige Kombination von Kommunikationsverbindungen bereitstellen, solange die genutzte Bandbreite der Kommunikationsverbindungen nicht die insgesamt verfügbare Bandbreite der physikalischen Verbindung übersteigt.

Wird eine neue Kommunikationsverbindung des Typs m angefordert zwischen einem ersten Knoten i und einem zweiten Knoten j (Endgeräte werden auch als Knoten bezeichnet), so kann die angeforderte Kommunikationsverbindung, wie oben dargestellt, entweder akzeptiert oder zurückgewiesen werden.

Wird die Kommunikationsverbindungen akzeptiert, so wird eine Route aus einer Menge vorgegebener Routen ausgewählt. Diese Auswahl wird als Routing bezeichnet. Im Rahmen der Kommunikationsverbindung vom Typ m werden b(m) Bandbreiteneinheiten für jede physikalische Verbindung entlang der ausgewählten Route für die Verbindungsdauer benutzt.

Somit kann im Rahmen der Zugangskontrolle (Call-Admission-Control) eine Route innerhalb des Kommunikationsnetzes 300 nur ausgewählt werden, wenn die ausgewählte Route ausreichend Bandbreite zur Verfügung hat.

Ziel der Zugangskontrolle und des Routings ist es, einen langfristigen Gewinn, der durch Akzeptanz der angeforderten Verbindungen erhalten wird, zu maximieren.

Das technische System Kommunikationsnetz 300 befindet sich zu einem Zeitpunkt t in einem Zustand x_t, welcher durch eine Liste von Routen über bestehende Verbindungen beschrieben wird, durch welche Listen angezeigt ist, wie viele Verbindungen welchen Diensttyps zu dem Zeitpunkt t die jeweilige Route verwenden.

Ereignisse w, durch die ein Zustand x_t in einen Folgezustand x_{t+1} überführt werden könnte, sind das Ankommen neuer Verbindungsanforderungsnachrichten oder auch das Beenden einer in dem Kommunikationsnetz 300 bestehenden Verbindung.

Eine Aktion a_t zu einem Zeitpunkt t aufgrund einer Verbindungsanforderung ist in diesem Ausführungsbeispiel die Ent-

scheidung, ob eine Verbindungsanforderung akzeptiert oder zurückgewiesen werden soll und, falls die Verbindung akzeptiert wird, die Auswahl der Route durch das Kommunikationsnetz 300.

Ziel ist die Ermittlung einer Folge von Aktionen, d.h. anschaulich das Lernen einer Strategie mit Aktionen zu einem Zustand x_t zu bestimmen derart, daß folgende Vorschrift maximiert wird:

$$E\left(\sum_{k=0}^{\infty} e^{-\beta t_k} \cdot g\left(x_{t_k}, \omega_k, a_{t_k}\right)\right), \qquad (20)$$

wobei mit

- $E\{.\}$ ein Erwartungswert,
- t_k ein Zeitpunkt, zu dem ein k-tes Ereignis erfolgt,
- $g\left(x_{t_k}, \omega_k, a_{t_k}\right)$ der Gewinn, der mit dem k-ten Ereignis verbunden ist, und
- β ein Verringerungsfaktor, der einen sofortigen Gewinn wertvoller bewertet als ein Gewinn in ferner in der Zukunft liegenden Zeitpunkten,

bezeichnet wird.

Unterschiedliche Realisierungen einer Strategie führen üblicherweise zu unterschiedlichen Gesamtgewinnen G:

$$G = \sum_{k=0}^{\infty} e^{-\beta t_k} \cdot g\left(x_{t_k}, \omega_k, a_{t_k}\right). \qquad (21)$$

Ziel ist die Maximierung des Erwartungswerts des Gesamtgewinns G gemäß folgender Vorschrift J:

$$J = E\left\{\sum_{k=0}^{\infty} e^{-\beta t_k} \cdot g\left(x_{t_k}, \omega_k, a_{t_k}\right)\right\}, \qquad (22)$$

wobei ein Risiko, daß der Gesamtgewinn G einer speziellen Realisierung einer Zugangskontrolle und einer Routing-Strategie unter den Erwartungswert sinkt, einstellbar ist.

Zur Durchführung der Zugriffskontrolle sowie zum Routing wird das TD(λ)-Lernverfahren eingesetzt.

Es wird folgende Zielfunktion im Rahmen dieses Ausführungsbeispiels eingesetzt:

$$J^*(x_t) = E_\tau\left\{e^{-\beta\tau}\right\}E_\omega\left\{\max_{a \in A}\left[g(x_t, \omega_t, a) + J^*(x_{t+1})\right]\right\}, \qquad (23)$$

wobei mit

- A ein Aktionsraum mit einer vorgegebenen Anzahl Aktionen, die in einem Zustand x_t jeweils zur Verfügung stehen,
- τ ein erster Zeitpunkt, zudem ein erstes Ereignis ω erfolgt,
- x_{t+1} ein Folgezustand des Systems,

bezeichnet wird.

Ein approximierter Wert des Zielwerts $J^*(x_t)$ wird unter Verwendung eines Funktionsapproximators 400 (vgl. **Fig.4**) unter Verwendung von Trainingsdaten gelernt und gespeichert.

Trainingsdaten sind zuvor in dem Kommunikationsnetz 300 gemessene Daten über das Verhalten des Kommunikationsnetzes 300 bei ankommenden Verbindungsanforderungen 304 sowie bei Beendigung von Nachrichten. Diese zeitliche Folge von Zuständen wird gespeichert und mit diesen Trainingsdaten wird der Funktionsapproximator 400 gemäß dem im weiteren beschriebenen Lernverfahren trainiert.

Als Eingangsgröße des Funktionsapproximators 400 dienen für jeden Eingang 401, 402, 403 des Funktionsapproximators 400 jeweils eine Anzahl von Verbindungen jeweils eines Diensttyps m auf einer Route des Kommunikationsnetzes 300. Diese sind in **Fig.4** durch Blöcke 404, 405, 406 symbolisch dargestellt.

Ausgangsgröße des Funktionsapproximators 400 ist ein approximierter Zielwert $\tilde{J}$ des Zielwerts J^*.

Eine detailliertere Darstellung des Funktionsapproximators 500, welcher in diesem Fall mehrere Teil-Funktionsapproximatoren 510, 520 des Funktionsapproximators 500 aufweist, zeigt **Fig.5**. Eine Ausgangsgröße ist der approximierte Zielwert $\tilde{J}$, der gemäß folgender Vorschrift gebildet wird:

$$\tilde{J}(x_t, \Theta) = \sum_{l=1}^{L} \tilde{J}^{(l)}\left(x_t^{(l)}, \Theta_t^{(l)}\right). \qquad (24)$$

Die Eingangsgrößen der Teilfunktionsapproximatoren 510, 520, die an Eingängen 511, 512, 513 des ersten Teil-Funktionsapproximators 510 bzw. an Eingängen 521, 522 und 523 des zweiten Teilfunktionsapproximators 520 anliegen, sind wiederum jeweils eine Anzahl von Diensttypen eines Typs m jeweils in einer physikalischen Verbindung r, symbolisiert durch Blöcke 514, 515, 516 für den ersten Teil-Funktionsapproximator und 524, 525 und 526 für den zweiten Teil-Funktionsapproximator 520.

Teilausgangsgrößen 530, 531, 532, 533 werden einer Addiereinheit 540 zugeführt und als Ausgangsgröße der Addiereinheit wird die approximierte Zielgröße $\tilde{J}$ gebildet.

Angenommen, das Kommunikationsnetz 300 befindet sich in dem Zustand x_{t_k} und eine Anforderungsnachricht, mit der ein Diensttyp m der Klasse m für eine Verbindung zwischen zwei

Knoten i, j angefordert wird, gelangt zu der ersten Verbindungseinheit 301a.

Mit R(i, j) wird eine Liste erlaubter Routen zwischen den Knoten i und j bezeichnet und mit

$$\tilde{R}\left(i, j, x_{t_k}\right) \subset R(i, j) \tag{25}$$

wird eine Liste aller möglichen Routen als Teilmenge der Routen R(i, j) bezeichnet, die hinsichtlich der verfügbaren und angeforderten Bandbreite eine mögliche Verbindung realisieren könnten.

Für jede mögliche Route r, $r \in \tilde{R}\left(i, j, x_{t_k}\right)$ wird ein Folgezustand $x_{t_k+1}\left(x_{t_k}, \omega_k, r\right)$ ermittelt, der daraus resultiert, daß die Verbindungsanforderung 304 akzeptiert wird und die Verbindung auf der Route r dem anfordernden ersten Endgerät 303 zur Verfügung gestellt wird.

Dies ist in **Fig.1** als zweiter Schritt (Schritt 102) dargestellt, wobei in einem ersten Schritt (Schritt 101) jeweils der Zustand des Systems sowie das jeweilige Ereignis festgestellt werden.

Es wird in einem dritten Schritt (Schritt 103) eine auszuwählende Route r* gemäß folgender Vorschrift ermittelt:

$$r^* = \arg \max_{r \in \tilde{R}\left(i, j, x_{t_k}\right)} \tilde{J}\left(x_{t_k+1}\left(x_{t_k}, \omega_k, r\right), \Theta_t\right). \tag{26}$$

In einem weiteren Schritt (Schritt 104) wird überprüft, ob folgende Vorschrift erfüllt ist:

$$c(m) + \tilde{J}\left(x_{t_k+1}\left(x_{t_k}, \omega_k, r^*\right), \Theta_t\right) < \tilde{J}\left(x_{t_k}, \Theta_t\right). \tag{27}$$

Ist dies der Fall, so wird die Verbindungsanforderung 304 zurückgewiesen (Schritt 105), sonst wird die Verbindung akzeptiert und entlang der ausgewählten Route r* zu dem Knoten j „durchgeschaltet" (Schritt 106).

In einem Parametervektor Θ sind jeweils für einen Zeitpunkt t Gewichte des Funktionsapproximators 400, 500 gespeichert, die im Rahmen des TD(λ)-Lernverfahrens an die Trainingsdaten adaptiert werden, so daß eine optimierte Zugangskontrolle und ein optimiertes Routing erreicht wird.

Während der Trainingsphase werden die Gewichtsparameter an die dem Funktionsapproximator angelegten Trainingsdaten angepaßt.

Es wird ein Risikoparameter κ definiert, mit dem ein gewünschtes Risiko, welches durch eine Folge von Aktionen und Zuständen hinsichtlich eines vorgegebenen Zustands des Systems aufweist, einstellbar ist, gemäß folgenden Vorschriften:

$-1 \leq \kappa < 0$: risikoreiches Lernen,

$\kappa = 0$: hinsichtlich des Risikos ein neutrales Lernen,

$0 < \kappa < 1$: ein risiko-vermeidendes Lernen,

$\kappa = 1$: „Worst-Case"-Lernen.

Fernen wird im Rahmen des Lernverfahrens ein vorgebbarer Parameter $0 \leq \lambda \leq 1$ und eine Schrittweitenfolge γ_k vorgegeben.

Die Gewichtswerte des Gewichtsvektors Θ werden aufgrund jedes Ereignisses ω_{t_k} gemäß folgender Adaptionsvorschrift an die Trainingsdaten angepaßt:

$$\Theta_k = \Theta_{k-1} + \gamma_k \aleph^{\kappa}(d_k) z_t , \qquad (28)$$

wobei

$$d_k = e^{-\beta(t_k - t_{k-1})}\left(g\left(x_{t_k}, \omega_k, a_{t_k}\right) + \tilde{J}\left(x_{t_k}, \Theta_{k-1}\right)\right) - \tilde{J}\left(x_{t_{k-1}}, \Theta_{k-1}\right) \tag{29}$$

$$z_t = \lambda e^{-\beta(t_{k-1} - t_{k-2})} z_{t-1} + \nabla_\Theta \tilde{J}\left(x_{t_{k-1}}, \Theta_{k-1}\right), \tag{30}$$

und

$$\aleph^\kappa(\xi) = \left(1 - \kappa \operatorname{sign}(\xi)\right)\xi . \tag{31}$$

Es wird angenommen: $Z_{-1} = 0$.

Die Funktion

$$g\left(x_{t_k}, \omega_k, a_{t_k}\right) \tag{32}$$

bezeichnet den sofortigen Gewinn gemäß folgender Vorschrift:

$$g\left(x_{t_k}, \omega_k, a_{t_k}\right) = \begin{cases} c(m) & \text{wenn } \omega_{t_k} \text{ ist eine Diens tan forderung eines Diensttyps m und die Verbindung wird akzeptiert} \\ 0 & \text{sonst} \end{cases} \tag{33}$$

Es wird also, wie oben beschrieben, eine Folge von Aktionen ermittelt, hinsichtlich einer Verbindungsanforderung, so daß eine Verbindungsanforderung aufgrund einer Aktion entweder zurückgewiesen oder akzeptiert wird. Die Ermittlung erfolgt unter Berücksichtigung einer Optimierungsfunktion, in der das Risiko mittels eines Risikokontrollparameters $\kappa \in [-1; 1]$ variabel einstellbar ist.

Zweites Ausführungsbeispiel: Verkehrsleitsystem

Fig.6 zeigt eine Straße 600, die von Autos 601, 602, 603, 604, 605 und 606 befahren ist.

In die Straße 600 integrierte Leiterschleifen 610, 611 nehmen elektrische Signale in bekannter Weise auf und führen die elektrischen Signale 615, 616, einem Rechner 620 über eine Eingangs-/Ausgangsschnittstelle 621 zu. In einem mit der Eingangs-/Ausgangsschnittstelle 621 verbundenen Analog-/Digital-Wandler 622 werden die elektrischen Signale in eine Zeitreihe digitalisiert und in einem Speicher 623, der über einen Bus 624 mit dem Analog-/Digital-Wandler 622 und einem Prozessor 625 verbunden ist, gespeichert. Über die Eingangs-/Ausgangsschnittsstelle 621 werden einem Verkehrsleitsystem 650 Steuerungssignale 651 zugeführt, aus denen in dem Verkehrsleitsystem 650 eine vorgegebene Geschwindigkeitsvorgabe 652 einstellbar ist oder auch weitere Angaben von Verkehrsvorschriften, die über das Verkehrsleitsystem 650 Fahrern der Fahrzeuge 601, 602, 603, 604, 605 und 606 dargestellt werden.

Zur Verkehrsmodellierung werden in diesem Fall folgende lokale Zustandsgrößen verwendet:

- Verkehrsflußgeschwindigkeit v,
- Fahrzeugdichte ρ (ρ = Anzahl von Fahrzeugen pro Kilometer $\frac{Fz}{km}$),
- Verkehrsfluß q (q = Anzahl der Fahrzeuge pro Stunde $\frac{Fz}{h}$, ($q = v * \rho$)), und
- jeweils zu einem Zeitpunkt von dem Verkehrsleitsystem 650 angezeigte Geschwindigkeitsbegrenzungen 652.

Die lokalen Zustandsgrößen werden wie oben beschrieben unter Verwendung der Leiterschleifen 610, 611 gemessen.

Somit stellen diese Größen (v(t), ρ(t), q(t)) einen Zustand des technischen Systems "Verkehr" zu einem bestimmten Zeitpunkt t dar.

In diesem Ausführungsbeispiel ist somit das System ein Verkehrssystem, welches unter Verwendung des Verkehrsleitsystems 650 geregelt wird.

Als Verfahren des approximativen dynamischen Programmierens wird in diesem zweiten Ausführungsbeispiel ein erweitertes Q-Lernverfahren beschrieben.

Der Zustand x_t wird beschrieben durch einen Zustandsvektor

$$x(t) = (v(t), \rho(t), q(t)) . \qquad (34)$$

Die Aktion a_t bezeichnet die Geschwindikeitsbegrenzung 652, die zum Zeitpunkt t von dem Verkehrsleitsystem 650 angezeigt wird.

Der Gewinn $r(x_t, a_t, x_{t+1})$ beschreibt die Güte des Verkehrsflusses, der zwischen den Zeitpunkten t und t+1 von den Leiterschleifen 610 und 611 gemessen wurde. Im Rahmen dieses zweiten Ausführungsbeispiels bezeichnet $r(x_t, a_t, x_{t+1})$

- die mittlere Geschwindigkeit der Fahrzeuge im Zeitintervall [t, t + 1],

oder

- die Anzahl der Fahrzeuge, die im Zeitintervall [t, t + 1] die Leiterschleifen 610 und 611 passiert haben,

oder

- die Varianz der Fahrzeuggeschwindigkeiten im Zeitintervall [t, t + 1],

oder

- eine gewichtete Summe aus den obigen Größen.

Für jede mögliche Aktion a_t, d.h. für jede von dem Verkehrsleitsystem 650 anzeigbare Geschwindigkeitsbegrenzung, wird ein Wert der Optimierungsfunktion OFQ ermittelt, wobei jeweils ein geschätzter Wert der Optimierungsfunktion OFQ als neuronales Netz realisiert wird.

Aus diese Weise ergibt sich eine Menge von Bewertungsgrößen für die unterschiedlichen Aktionen a_t in dem Systemzustand x_t.

In einer Regelungsphase wird aus den möglichen Aktionen a_t, d.h. aus der Menge der von dem Verkehrsleitsystem 650 anzeigbaren Geschwindigkeitsbegrenzungen, diejenige Aktion a_t ausgewählt, für die in dem aktuellen Systemzustand x_t die maximale Bewertungsgröße OFQ ermittelt worden ist.

Die aus dem Q-Lernverfahren bekannte Adaptionsvorschrift zur Berechnung der Optimierungsfunktion OFQ wird gemäß diesem Ausführungsbeispiel um eine Risikokontrollfunktion $\aleph^{\kappa}(\cdot)$, die das Risiko berücksichtigt, erweitert.

Wiederum wird der Risikokontrollparameter κ gemäß der Strategie aus dem ersten Ausführungsbeispiel im Intervall von $[-1 \leq \kappa \leq 1]$ vorgegeben und repräsentiert das Risiko, das ein Benutzer im Rahmen der Anwendung hinsichtlich der zu bestimmenden Kontrollstrategie eingehen will.

Gemäß diesem Ausführungsbeispiel wird folgende Bewertungsfunktion OFQ verwendet:

$$OFQ = Q\left(x; w^a\right), \tag{35}$$

wobei mit

- $x = (v; \rho; q)$ ein Zustand des Verkehrssystems,
- a eine Geschwindigkeitsbegrenzung aus dem Aktionsraum A aller vom Verkehrsleitsystem 650 anzeigbaren Geschwindigkeitsbegrenzungen,
- w^a die zur Geschwindigkeitsbegrenzung a gehörigen Gewichte des neuronalen Netzes,

bezeichnet wird/werden.

Im Rahmen des Q-Lernens wird zur Ermittlung der optimalen Gewichte w^a den neuronalen Netzes folgender Adaptionsschritt ausgeführt:

$$w_{t+1}^{a_t} = w_t^{a_t} + \eta_t \cdot \aleph^{\kappa}(d_t) \cdot \nabla Q\left(x_t; w_t^{a_t}\right) \qquad (36)$$

mit der Abkürzung

$$d_t = r(x_t, a_t, x_{t+1}) + \gamma \max_{a \in A} Q\left(x_{t+1}, w_t^a\right) - Q\left(x_t, w_t^{a_t}\right) \qquad (37)$$

wobei mit

- x_t, x_{t+1} jeweils ein Zustand des Verkehrssystems gemäß Vorschrift (34),
- a_t eine Aktion, d.h. eine von dem Verkehrsleitsystem 650 anzeigbare Geschwindigkeitsbegrenzung,
- γ ein vorgebbarer Verringerungsfaktor,
- $w_t^{a_t}$ der zur Aktion a_t gehörige Gewichtsvektor vor dem Adaptionsschritt,
- $w_{t+1}^{a_t}$ der zur Aktion a_t gehörige Gewichtsvektor nach dem Adaptionsschritt,
- η_t (t = 1, ...) eine vorgebbare Schrittweitenfolge,

- $\kappa \in [-1; 1]$ ein Risikokontrollparameter,
- $\aleph^{\kappa}$ eine Risikokontrollfunktion $\aleph^{\kappa}(\xi) = (1 - \kappa \mathrm{sign}(\xi))\xi$,
- $\nabla Q(\cdot;\cdot)$ die Ableitung des neuronalen Netzes nach seinen Gewichten,
- $r(x_t, a_t, x_{t+1})$ ein Gewinn beim Zustandsübergang von dem Zustand x_t nach dem Folgezustand x_{t+1},

bezeichnet wird/werden.

Im Rahmen des Lernens kann aus den möglichen Aktionen a_t eine Aktion a_t zufällig gewählt werden. Hierbei ist es nicht erforderlich, die Aktion a_t, die zu der größten Bewertungsgröße geführt hat, zu wählen.

Die Adaption der Gewichte hat derart zu erfolgen, daß nicht nur eine im Erwartungswert der Optimierungsfunktion optimierte Regelung des Verkehrs erreicht wird, sondern auch eine Varianz der Regelungsergebnisse berücksichtigt wird.

Dies ist insbesondere vorteilhaft, da der Zustandsvektor x(t) das tatsächliche System Verkehr in einigen Aspekten nur unzureichend modelliert und es deshalb zu nicht erwarteten Störungen kommen kann. So hängt die Dynamik des Verkehrs und damit seiner Modellierung von weiteren Faktoren wie beispielsweise Wetter, Anteil an Lastkraftwagen auf der Straße, ein Anteil von Wohnmobilen, etc. ab, die nicht immer in den Meßgrößen des Zustandsvektors x(t) integriert sind. Zudem ist nicht immer sichergestellt, daß die Verkehrsteilnehmer sofort den neuen Geschwindigkeitsangaben gemäß dem Verkehrsleitsystem Folge leisten.

Eine Regelungsphase an dem realen System gemäß dem Verkehrsleitsystem vollzieht sich gemäß folgenden Schritten:

1. Das Messen des Zustandes x_t zum Zeitpunkt t erfolgt an verschiedenen Stellen des Verkehrssystems Verkehr und ergibt einen Zustandsvektor $x(t) = (v(t), \rho(t), q(t))$.

2. Für alle möglichen Aktionen a_t wird ein Wert der Optimierungsfunktion ermittelt und es wird diejenige Aktion a_t mit der höchsten Bewertung in der Optimierungsfunktion ausgewählt.

In diesem Dokument sind folgende Veröffentlichungen zitiert:

[1] R. Neuneier, Enhancing Q-Learning for Optimal Asset Allocation, Proceedings of the Neural Information Processing Systems, NIPS 1997

[2] R.S. Sutton, Learning to predict by the method of temporal differences, Machine Learning, 3:9-44, 1988

[3] M. Heger, Risk and Reinforcement Learning: Concepts and Dynamic Programming, ZKW Bericht Nr. 8/94, Zentrum für Kognitionswissenschaften, Universität Bremen, ISSN 0947-0204, Dezember 1994

[4] D.P. Bertsekas, Dynamic Programming and Optimal Control, Athena Scientific, Belmont, MA, 1995

Patentansprüche

1. Verfahren zur rechnergestützten Ermittlung einer Folge von Aktionen für ein System, welches Zustände aufweist, wobei ein Zustandsübergang zwischen zwei Zuständen aufgrund einer Aktion erfolgt,
bei dem die Ermittlung der Folge von Aktionen derart erfolgt, daß eine aus der Folge von Aktionen resultierende Folge von Zuständen hinsichtlich einer vorgegebenen Optimierungsfunktion optimiert ist, wobei die Optimierungsfunktion einen variablen Parameter enthält, mit dem ein Risiko, welches die resultierende Folge von Zuständen hinsichtlich eines vorgegebenen Zustandes des Systems aufweist, einstellbar ist.

2. Verfahren nach Anspruch 1,
bei dem zur Ermittlung ein Verfahren des approximativen Dynamischen Programmierens eingesetzt wird.

3. Verfahren nach Anspruch 2,
bei dem das Verfahren des approximativen Dynamischen Programmierens ein auf dem Q-Lernen basierendes Verfahren ist.

4. Verfahren nach Anspruch 3,
bei dem die Optimierungsfunktion OFQ im Rahmen des Q-Lernens gemäß folgender Vorschrift gebildet wird:

$$OFQ = Q\left(x; w^a\right),$$

wobei mit

- x ein Zustand in einem Zustandsraum X,
- a eine Aktion aus einem Aktionsraum A,
- w^a die zur Aktion a gehörigen Gewichte eines Funktionsapproximators

bezeichnet wird/werden, und bei dem die Gewichte des Funktionsapproximators gemäß folgender Vorschrift adaptiert werden:

$$w_{t+1}^{a_t} = w_t^{a_t} + \eta_t \cdot \aleph^{\kappa}(d_t) \cdot \nabla Q\left(x_t; w_t^{a_t}\right)$$

mit der Abkürzung

$$d_t = r(x_t, a_t, x_{t+1}) + \gamma \max_{a \in A} Q\left(x_{t+1}, w_t^{a}\right) - Q\left(x_t, w_t^{a_t}\right)$$

wobei mit

- x_t, x_{t+1} jeweils ein Zustand in dem Zustandsraum X,
- a_t eine Aktion aus einem Aktionsraum A,
- γ ein vorgebbarer Verringerungsfaktor,
- $w_t^{a_t}$ der zur Aktion a_t gehörige Gewichtsvektor vor dem Adaptionsschritt,
- $w_{t+1}^{a_t}$ der zur Aktion a_t gehörige Gewichtsvektor nach dem Adaptionsschritt,
- η_t (t = 1, ...) eine vorgebbare Schrittweitenfolge,
- $\kappa \in [-1; 1]$ ein Risikokontrollparameter,
- $\aleph^{\kappa}$ eine Risikokontrollfunktion $\aleph^{\kappa}(\xi) = (1 - \kappa \operatorname{sign}(\xi))\xi$,
- $\nabla Q(\cdot;\cdot)$ die Ableitung des Funktionsapproximators nach seinen Gewichten,
- $r(x_t, a_t, x_{t+1})$ ein Gewinn beim Zustandsübergang von dem Zustand x_t nach dem Folgezustand x_{t+1},

bezeichnet wird/werden.

5. Verfahren nach Anspruch 2,
bei dem das Verfahren des approximativen Dynamischen Programmierens ein auf dem TD(λ)-Lernen basierendes Verfahren ist.

6. Verfahren nach Anspruch 5,
bei dem die Optimierungsfunktion OFTD im Rahmen des TD(λ)-Lernens gemäß folgender Vorschrift gebildet wird:

$$OFTD = J(x; w)$$

wobei mit

- x ein Zustand in einem Zustandsraum X,
- a eine Aktion aus einem Aktionsraum A,
- w die Gewichte eines Funktionsapproximators

bezeichnet wird/werden, und bei dem die Gewichte des Funktionsapproximators gemäß folgender Vorschrift adaptiert werden:

$$w_{t+1} = w_t + \eta_t \cdot \aleph^{\kappa}(d_t) \cdot z_t$$

mit den Abkürzungen

$$d_t = r(w_t, a_t, x_{t+1}) + \gamma J(x_{t+1}; w_t) - J(x_t; w_t),$$

$$z_t = \lambda \cdot \gamma \cdot z_{t-1} + \nabla J(x_t; w_t),$$

$$z_{-1} = 0,$$

wobei mit

- x_t, x_{t+1} jeweils ein Zustand in dem Zustandsraum X,
- a_t eine Aktion aus einem Aktionsraum A,
- γ ein vorgebbarer Verringerungsfaktor,
- w_t der Gewichtsvektor vor dem Adaptionsschritt,
- w_{t+1} der Gewichtsvektor nach dem Adaptionsschritt,
- η_t $(t = 1, \ldots)$ eine vorgebbare Schrittweitenfolge,
- $\kappa \in [-1; 1]$ ein Risikokontrollparameter,
- $\aleph^{\kappa}$ eine Risikokontrollfunktion $\aleph^{\kappa}(\xi) = (1 - \kappa \operatorname{sign}(\xi))\xi$,
- $\nabla J(\cdot;\cdot)$ die Ableitung des Funktionsapproximators nach seinen Gewichten,
- $r(x_t, a_t, x_{t+1})$ ein Gewinn beim Zustandsübergang von dem Zustand x_t nach dem Folgezustand x_{t+1},

bezeichnet wird/werden.

7. Verfahren nach einem der Ansprüche 1 bis 6,
bei dem das System ein technisches System ist, von dem vor der Ermittlung Meßgrößen gemessen werden, die bei der Ermittlung der Folge von Aktionen verwendet werden.

8. Verfahren nach Anspruch 7,
bei dem gemäß der Folge von Aktionen das technische System gesteuert wird.

9. Verfahren nach Anspruch 7,
bei dem gemäß der Folge von Aktionen das technische System geregelt wird.

10. Verfahren nach einem der Ansprüche 1 bis 9,
bei dem das System als ein Markov-Entscheidungsproblem modelliert wird.

11. Verfahren nach einem der Ansprüche 1 bis 10,
eingesetzt in einem Verkehrsleitsystem.

12. Verfahren nach einem der Ansprüche 1 bis 10,
eingesetzt in einem Kommunikationssystem.

13. Verfahren nach einem der Ansprüche 1 bis 10,
eingesetzt zur Durchführung einer Zugangskontrolle in einem Kommunikationsnetz.

14. Verfahren nach einem der Ansprüche 1 bis 10,
eingesetzt zur Durchführung eines Routings in einem Kommunikationsnetz.

15. Anordnung zur Ermittlung einer Folge von Aktionen für ein System, welches Zustände aufweist, wobei ein Zustandsübergang zwischen zwei Zuständen aufgrund einer Aktion erfolgt,

mit einem Prozessor, der derart eingerichtet ist, daß die Ermittlung der Folge von Aktionen derart erfolgen kann, daß eine aus der Folge von Aktionen resultierende Folge von Zuständen hinsichtlich einer vorgegebenen Optimierungsfunktion optimiert ist, wobei die Optimierungsfunktion einen variablen Parameter enthält, mit dem ein Risiko, welches die resultierende Folge von Zuständen hinsichtlich eines vorgegebenen Zustandes des Systems aufweist, einstellbar ist.

16. Anordnung nach Anspruch 15,
eingesetzt zur Steuerung eines technischen Systems.

17. Anordnung nach Anspruch 15,
eingesetzt zur Regelung eines technischen Systems.

18. Anordnung nach Anspruch 15,
eingesetzt in einem Verkehrsleitsystem.

19. Anordnung nach Anspruch 15,
eingesetzt in einem Kommunikationssystem.

20. Anordnung nach Anspruch 15,
eingesetzt zur Durchführung einer Zugangskontrolle in einem Kommunikationsnetz.

21. Anordnung nach Anspruch 15,
eingesetzt zur Durchführung eines Routings in einem Kommunikationsnetz.

FIG 1

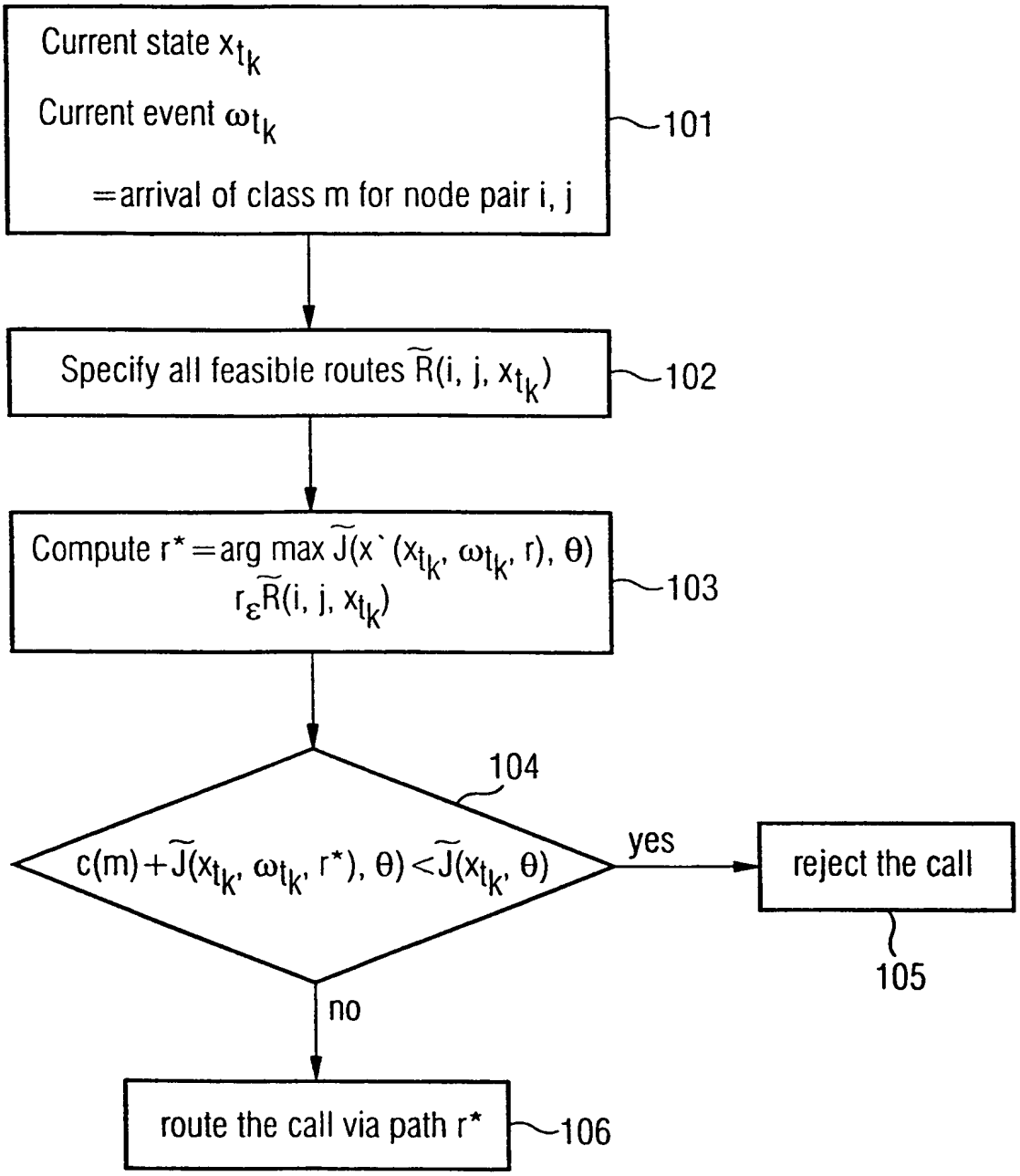

Current state x_{t_k}
Current event ω_{t_k}
=arrival of class m for node pair i, j
101
Specify all feasible routes $\widetilde{R}(i, j, x_{t_k})$
102
Compute $r^* = \arg\max_{r \varepsilon \widetilde{R}(i, j, x_{t_k})} \widetilde{J}(x'(x_{t_k}, \omega_{t_k}, r), \theta)$
103
104
$c(m) + \widetilde{J}(x_{t_k}, \omega_{t_k}, r^*), \theta) < \widetilde{J}(x_{t_k}, \theta)$
yes
reject the call
105
no
route the call via path r*
106

FIG 2

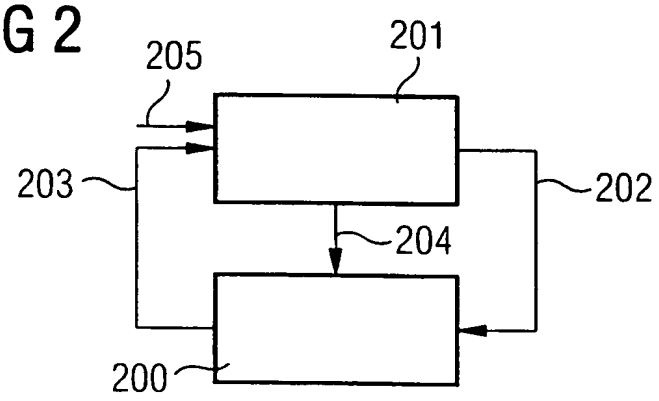

205
201
203
202
204
200

FIG3

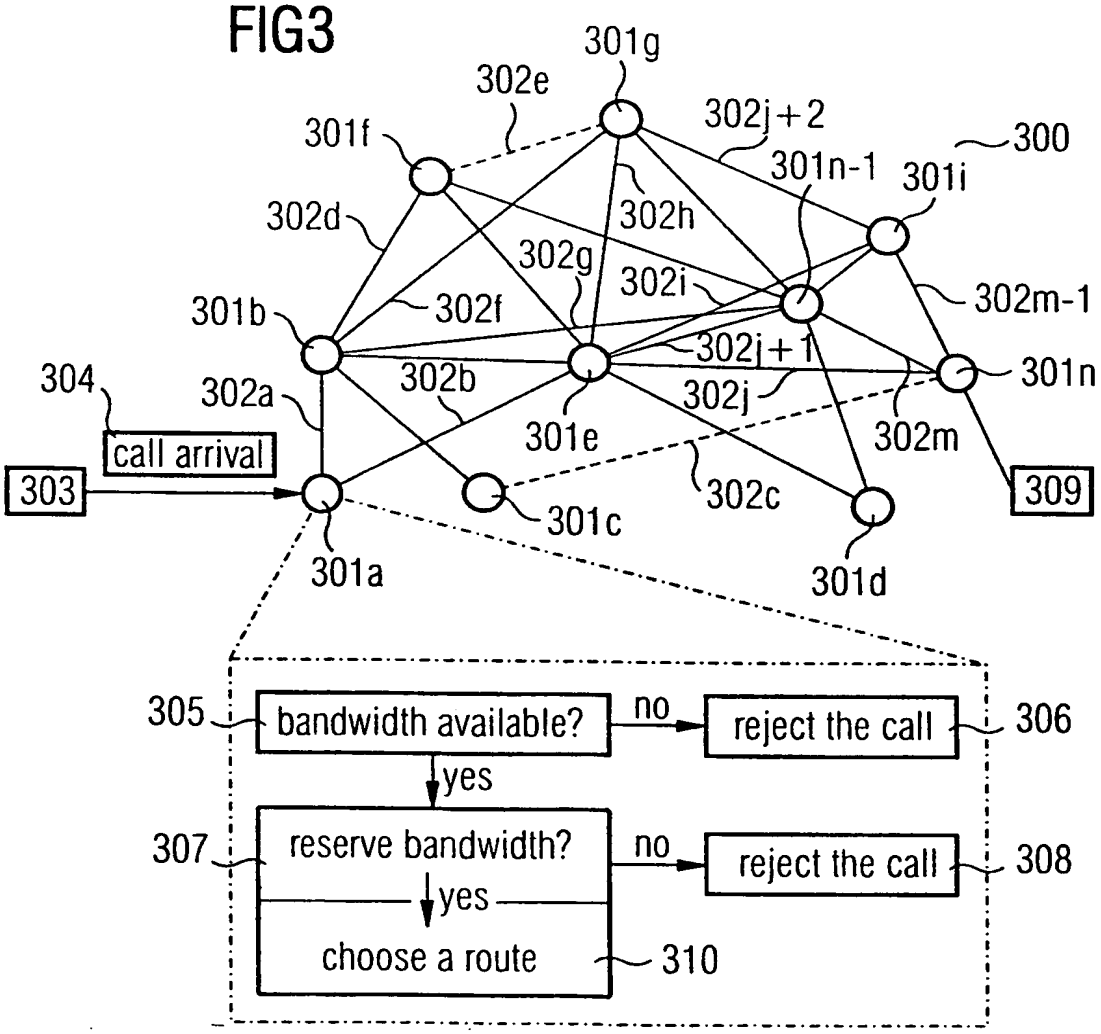

301g
302e
302j+2
301f
300
301n-1
301i
302d
302h
302g
302i
302m-1
301b
302f
302j+1
301n
304
302a
302b
302j
call arrival
301e
302m
303
302c
309
301c
301a
301d
305
bandwidth available?
no
reject the call
306
yes
307
reserve bandwidth?
no
reject the call
308
yes
choose a route
310

FIG 4

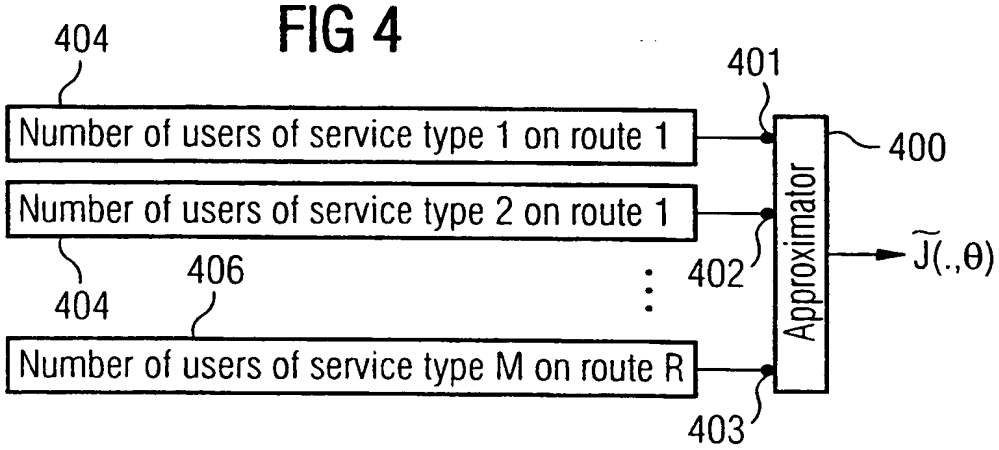

404
401
Number of users of service type 1 on route 1
Number of users of service type 2 on route 1
404
406
Number of users of service type M on route R
402
403
Approximator
400
$\tilde{J}(.,\theta)$

FIG 5

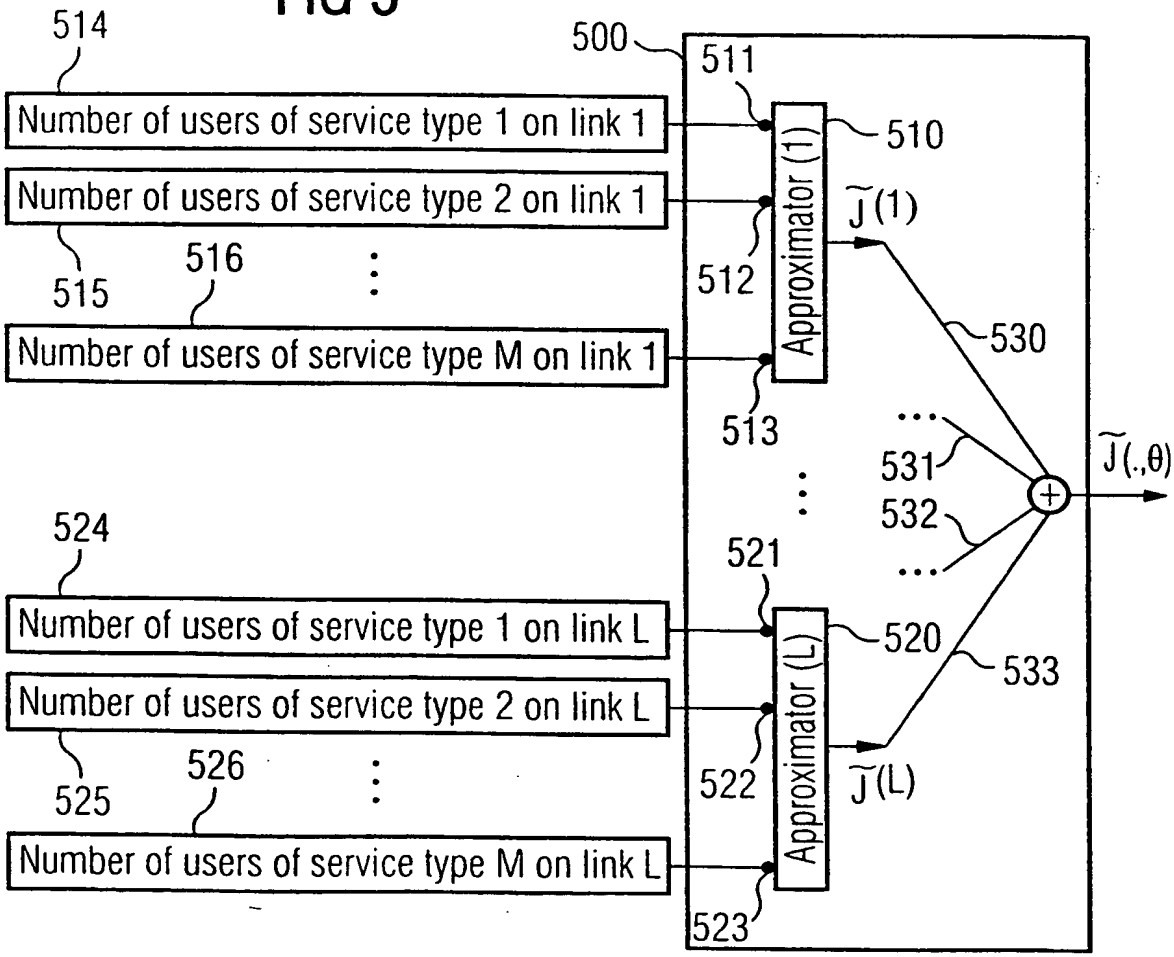
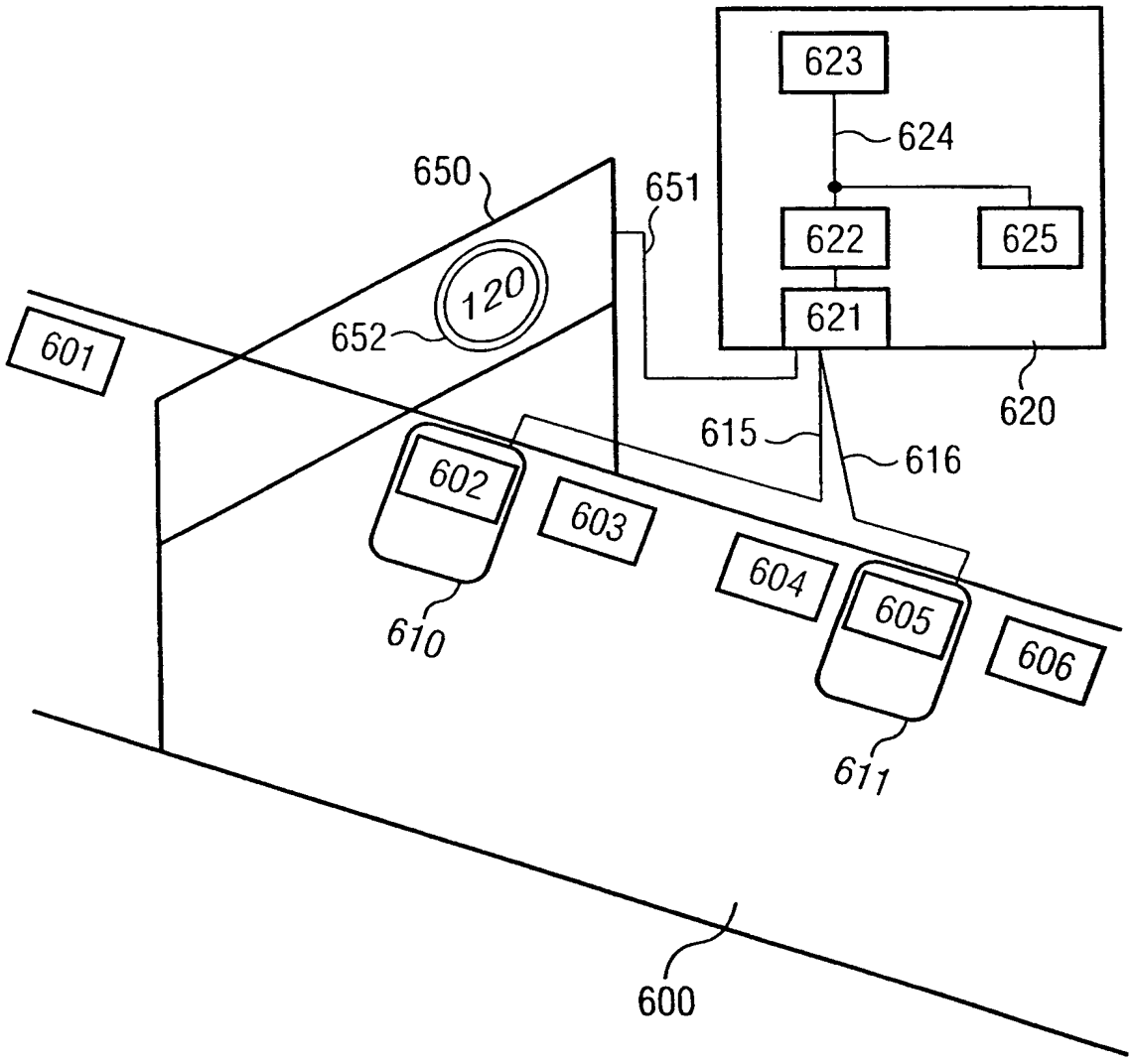

514
500
511
Number of users of service type 1 on link 1
Number of users of service type 2 on link 1
515
516
Number of users of service type M on link 1
512
513
Approximator (1)
510
$\tilde{J}(1)$
530
531
532
533
$\tilde{J}(.,\theta)$
524
521
Number of users of service type 1 on link L
Number of users of service type 2 on link L
525
526
Number of users of service type M on link L
522
523
Approximator (L)
520
$\tilde{J}(L)$

FIG 6


623
624
622
625
621
650
651
120
652
601
620
615
616
602
603
604
605
606
610
611
600

(12) NACH DEM VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS (PCT) VERÖFFENTLICHTE INTERNATIONALE ANMELDUNG

(19) Weltorganisation für geistiges Eigentum
Internationales Büro

(43) Internationales Veröffentlichungsdatum
30. März 2000 (30.03.2000)

PCT

(10) Internationale Veröffentlichungsnummer
WO 00/17811 A3

(51) Internationale Patentklassifikation[7]: G06F 17/60, G05B 13/02

(21) Internationales Aktenzeichen: PCT/DE99/02846

(22) Internationales Anmeldedatum:
8. September 1999 (08.09.1999)

(25) Einreichungssprache: Deutsch

(26) Veröffentlichungssprache: Deutsch

(30) Angaben zur Priorität:
198 43 620.3 23. September 1998 (23.09.1998) DE

(71) Anmelder (*für alle Bestimmungsstaaten mit Ausnahme von US*): **SIEMENS AKTIENGESELLSCHAFT** [DE/DE]; Wittelsbacherplatz 2, D-80333 München (DE).

(72) Erfinder; und

(75) Erfinder/Anmelder (*nur für US*): **NEUNEIER, Ralf** [DE/DE]; Gravelottestr. 3, D-81667 München (DE). **MIHATSCH, Oliver** [DE/DE]; Schulstr. 31, D-80634 München (DE).

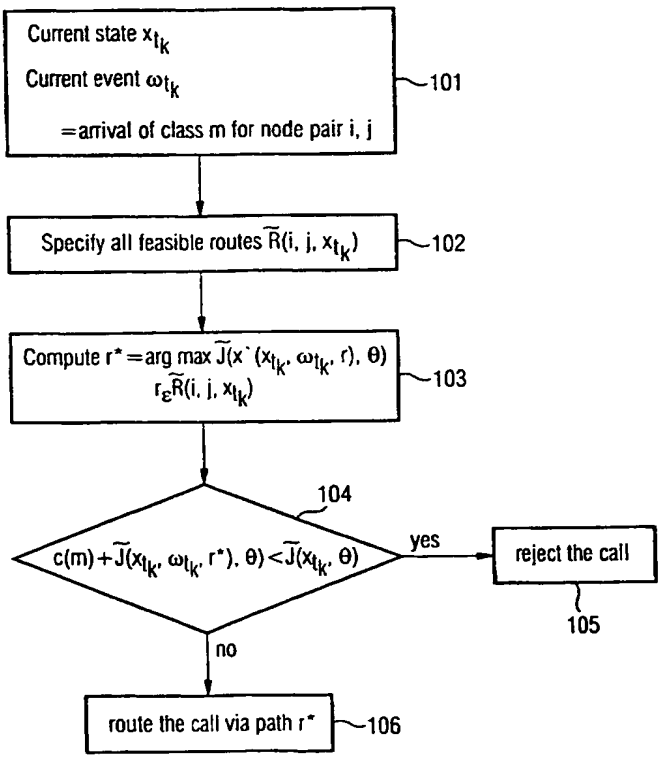
(74) Gemeinsamer Vertreter: **SIEMENS AKTIENGESELLSCHAFT**; Postfach 22 16 34, D-80506 München (DE).

(81) Bestimmungsstaaten (*national*): JP, US.

[Fortsetzung auf der nächsten Seite]

(54) Title: METHOD AND CONFIGURATION FOR DETERMINING A SEQUENCE OF ACTIONS FOR A SYSTEM WHICH COMPRISES STATUSES, WHEREBY A STATUS TRANSITION ENSUES BETWEEN TWO STATUSES AS A RESULT OF AN ACTION

(54) Bezeichnung: VERFAHREN UND ANORDNUNG ZUR ERMITTLUNG EINER FOLGE VON AKTIONEN FÜR EIN SYSTEM, WELCHES ZUSTÄNDE AUFWEIST, WOBEI EIN ZUSTANDSÜBERGANG ZWISCHEN ZWEI ZUSTÄNDEN AUFGRUND EINER AKTION ERFOLGT



(57) Abstract: The determination of a sequence of actions ensues in such a way that a sequence of statuses resulting from the sequence of actions is optimized with regard to a predetermined optimization function. The optimization function includes a variable parameter with which a risk can be set. Said risk comprises the resulting sequence of statuses with regard to a predetermined status of the system.

(57) Zusammenfassung: Die Ermittlung der Folge von Aktionen erfolgt derart, daß eine aus der Folge von Aktionen resultierende Folge von Zuständen hinsichtlich einer vorgegebenen Optimierungsfunktion optimiert ist. Die Optimierungsfunktion enthält einen variablen Parameter, mit dem ein Risiko einstellbar ist, welches Risiko die resultierende Folge von Zuständen hinsichtlich eines vorgegebenen Zustandes des Systems aufweist.

WO 00/17811 A3

(84) **Bestimmungsstaaten** *(regional)*: europäisches Patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Veröffentlicht:
— *Mit internationalem Recherchenbericht.*

(88) **Veröffentlichungsdatum des internationalen Recherchenberichts:** 7. Dezember 2000

Zur Erklärung der Zweibuchstaben-Codes, und der anderen Abkürzungen wird auf die Erklärungen ("Guidance Notes on Codes and Abbreviations") am Anfang jeder regulären Ausgabe der PCT-Gazette verwiesen.

INTERNATIONAL SEARCH REPORT

Intern. al Application No
PCT/DE 99/02846

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F17/60 G05B13/02

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06F G05B

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, INSPEC, PAJ, COMPENDEX, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	NEUNEIER R: "Enhancing Q-Learning for Optimal Asset Allocation" ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 10-PROCEEDINGS OF THE 1997 CONFERENCE, 1997, pages 936-942, XP002144732	1-3, 15-21
A	the whole document	4-14
Y	GUTJAHR W J: "Failure Risk Estimation via Markov Software Usage Models" PROCEEDINGS OF THE 15TH INTERNATIONAL CONFERENCE ON COMPUTER SAFETY, RELIABILITY AND SECURITY, SAFECOMP 96, 23 - 25 October 1996, pages 183-192, XP000921432	1-3, 15-21
A	page 185, paragraph 3	4-14

-/--

[X] Further documents are listed in the continuation of box C. [ ] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
11 August 2000	24/08/2000

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Skulikaris, I

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Intern. al Application No
PCT/DE 99/02846

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	AOKI Y: "EVALUATION AND OPTIMIZATION OF ENVIRONMENTAL PLANNING UNDER THE RISK-AVERSION OF NON-REPAIRABLE DAMAGE" ENVIRONMENTAL SYSTEMS PLANNING CONFERENCE, 1 - 5 August 1977, pages 847-852, XP000922719 the whole document	1-21

INTERNATIONALER RECHERCHENBERICHT

Intern ales Aktenzeichen
PCT/DE 99/02846

A. KLASSIFIZIERUNG DES ANMELDUNGSGEGENSTANDES
IPK 7 G06F17/60 G05B13/02

Nach der Internationalen Patentklassifikation (IPK) oder nach der nationalen Klassifikation und der IPK

B. RECHERCHIERTE GEBIETE

Recherchierter Mindestprüfstoff (Klassifikationssystem und Klassifikationssymbole)
IPK 7 G06F G05B

Recherchierte aber nicht zum Mindestprüfstoff gehörende Veröffentlichungen, soweit diese unter die recherchierten Gebiete fallen

Während der internationalen Recherche konsultierte elektronische Datenbank (Name der Datenbank und evtl. verwendete Suchbegriffe)
EPO-Internal, INSPEC, PAJ, COMPENDEX, WPI Data

C. ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie*	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
Y	NEUNEIER R: "Enhancing Q-Learning for Optimal Asset Allocation" ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 10-PROCEEDINGS OF THE 1997 CONFERENCE, 1997, Seiten 936-942, XP002144732	1-3, 15-21
A	das ganze Dokument	4-14
Y	GUTJAHR W J: "Failure Risk Estimation via Markov Software Usage Models" PROCEEDINGS OF THE 15TH INTERNATIONAL CONFERENCE ON COMPUTER SAFETY, RELIABILITY AND SECURITY, SAFECOMP 96, 23. - 25. Oktober 1996, Seiten 183-192, XP000921432	1-3, 15-21
A	Seite 185, Absatz 3	4-14

-/--

[X] Weitere Veröffentlichungen sind der Fortsetzung von Feld C zu entnehmen

[ ] Siehe Anhang Patentfamilie

* Besondere Kategorien von angegebenen Veröffentlichungen :
"A" Veröffentlichung, die den allgemeinen Stand der Technik definiert, aber nicht als besonders bedeutsam anzusehen ist
"E" älteres Dokument, das jedoch erst am oder nach dem internationalen Anmeldedatum veröffentlicht worden ist
"L" Veröffentlichung, die geeignet ist, einen Prioritätsanspruch zweifelhaft erscheinen zu lassen, oder durch die das Veröffentlichungsdatum einer anderen im Recherchenbericht genannten Veröffentlichung belegt werden soll oder die aus einem anderen besonderen Grund angegeben ist (wie ausgeführt)
"O" Veröffentlichung, die sich auf eine mündliche Offenbarung, eine Benutzung, eine Ausstellung oder andere Maßnahmen bezieht
"P" Veröffentlichung, die vor dem internationalen Anmeldedatum, aber nach dem beanspruchten Prioritätsdatum veröffentlicht worden ist
"T" Spätere Veröffentlichung, die nach dem internationalen Anmeldedatum oder dem Prioritätsdatum veröffentlicht worden ist und mit der Anmeldung nicht kollidiert, sondern nur zum Verständnis des der Erfindung zugrundeliegenden Prinzips oder der ihr zugrundeliegenden Theorie angegeben ist
"X" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann allein aufgrund dieser Veröffentlichung nicht als neu oder auf erfinderischer Tätigkeit beruhend betrachtet werden
"Y" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann nicht als auf erfinderischer Tätigkeit beruhend betrachtet werden, wenn die Veröffentlichung mit einer oder mehreren anderen Veröffentlichungen dieser Kategorie in Verbindung gebracht wird und diese Verbindung für einen Fachmann naheliegend ist
"&" Veröffentlichung, die Mitglied derselben Patentfamilie ist

Datum des Abschlusses der internationalen Recherche	Absendedatum des internationalen Recherchenberichts
11. August 2000	24/08/2000

Name und Postanschrift der Internationalen Recherchenbehörde	Bevollmächtigter Bediensteter
Europäisches Patentamt, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Tx. 31 651 epo nl, Fax: (+31–70) 340–3016	Skulikaris, I

Formblatt PCT/ISA/210 (Blatt 2) (Juli 1992)

C.(Fortsetzung) ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie°	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
A	AOKI Y: "EVALUATION AND OPTIMIZATION OF ENVIRONMENTAL PLANNING UNDER THE RISK-AVERSION OF NON-REPAIRABLE DAMAGE" ENVIRONMENTAL SYSTEMS PLANNING CONFERENCE, 1. - 5. August 1977, Seiten 847-852, XP000922719 das ganze Dokument	1-21